A STUDY DATA MINING TECHNIQUE FOR BREAST CANCER PREDICTIONUSING FUZZY THEORY

A Synopsis

Submitted to IIMT University, Meerut

In partial fulfilment for the award of the Degree of Doctor of Philosophy

(Ms. Parul Bhatnagar)

Enrollment No- IIMTU/2166342 *Registration No.21210963*

Under the Guidance of

Dr. BHUPENDRA KUMAR
School of Computer Science & Applications
IIMT University, Meerut

Department of

School of Computer Science & Applications

IIMT University, Meerut

IIMT University, Meeru

1. Introduction

By 2020, 2.35 million women will be diagnosed with breast cancer, and 685,000 will die from it, according to the WHO. The deadliest female malignancy is breast cancer. Breast cancer deaths are mostly caused by metastasis and recurrence. Despite initial treatment, metastatic recurrence of breast cancer could occur somewhere around weeks to years later (Magboo et al., 2021).

Furthermore, data organization begins immediately. Most mammary glands and ducts could develop breast cancer. Breast cancer cells in circulation could spread (Ghiasi et al., 2021) so that Fuzzy System plays a vital role in determining the logic and characteristics of the system. The four main parts of a fuzzy system are shown in Figure 1. These are the fuzzifier, the inference engine, the de-fuzzifier, and the fuzzy base or knowledge base (Idris et al., 2021).

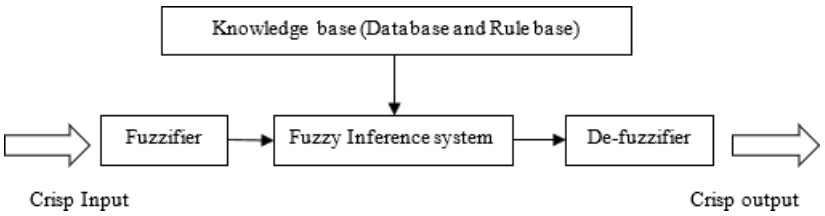

Figure 1: Features of the fuzzy system (Idris et al., 2021).

Age, gender, genetics, reproductive history, and hormonal condition increase breast cancer risk (Jevtic et al., 2010). If breast cancer is not treated early, metastasis—the spread of cancer—can occur. Lung, liver, bone, and brain cancers are possible. Stroke hazards include blood clots, shattered bones, and spinal cord compression. Data-driven prediction technology may improve cancer prognosis. Data mining has been effective in numerous healthcare fields of study, including cancer treatment (Walczak et al., 2018). The healthcare industry needs several data mining approaches to make sense of enormous data sets and get useful insights. Data mining in healthcare is expected to save costs by 30%. Hospitals and clinics are adopting EHR systems. Data mining may improve medical care since healthcare providers have more patient records than ever before (Kourou et al., 2015).

Data mining techniques for breast cancer

Academics struggle to identify and predict breast cancer. Machine learning and data mining have improved breast cancer diagnosis, prediction, and the overall method. Data mining is replacing

decision help in early breast cancer detection (Gupta et al., 2011). . Learning strategies including supervised, unsupervised, and semi-supervised learning are all used in data mining. Instances are classified into categories based on their values of the target attributes or the class attributes (Lashari et al., 2014).

Aims and Objectives

- To provide a practical approach to dealing with tumor formation using the existing cancer dataset.
- To evaluate the efficacy of various classifiers, such as DT, SVM, K-NN, and NB, and draw conclusions from the data.
- Using the established ensemble approach, obtain the highest possible fitness score.

2. Review of literature

The following study uses fuzzy theory to predict breast cancer through data mining. Below is a collection of researchers' interpretations.

Lanbaran et al., (2021) demonstrated early breast cancer detection requires a prompt medical examination. Cancer treatment cost varies. This study classified breast cancer using machine learning and sensitivity intuitionistic fuzzy-rough set. Choosing features reduces extraneous components. Data reduction improves massive data. Rough set theory dissecting real-value datasets could minimize features but lose information. The tolerance-based technique was incorrect. To show how machine learning and feature selection work which is the greatest way to speed up the process and improve accuracy.

Chaudhari et al., (2020) analysis the breast cancer is the leading cause of death among women worldwide, and the epidemic of cancer that is now unfolding is having a disastrous effect on humanity as a whole. The recurrence of symptoms is a symptom of this condition. Errors of type 1 and type 2 in recurrence forecasting are discussed, and a unified approach is anticipated. After removing inconsistent variables from the database, the results reveal an increase in accuracy when the focus is narrowed to just the most crucial ones.

Akinnuwesi et al., (2020) stated that breast cancer (BC) is one of the most common cancers in women globally, despite its unknown cause. SVM and PCA were the first analytical approaches to quickly find BCA in research. PCA extracted structures in the first step of preprocessing,

however the second stage decreased feature collection. After pre-processing multi-predicted data, SVM predicted breast cancer risk and diagnosis. It detects early cancer and assesses risk with 97% accuracy, 95% specificity, and 100% sensitivity (BC-RAED). BC-RAED uses technological markers to detect cancer and benign disorders.

Abdar et. Al., (2020) suggested that breast cancer in women is one of the most common forms of the disease. This inquiry explores the potential of data mining and machine learning-based methods for automating the identification of breast cancer in females. Thus, the BayesNet approach was more effective than the Naive Bayes algorithm. With K=10, the BayesNet approach is determined to have the highest accuracy (95.25 %), whereas the Naive Bayes algorithm has the greatest results (93.32 %) when K=3 (93.32 percent).

Mojrian et al., (2020) suggested that mammography is among the most efficient ways for detecting breast cancer in the laboratories, and one of the largest and most popular approaches. An Extreme learning machine-Radial basis function (ELM-RBF) multilayer fuzzy expert system has been reported for breath cancer detection using the Wisconsin dataset. Using the Wisconsin dataset as guidance, it combined an ELM instructional with an RBF kernel to create a multilayered fuzzy expert system for breast cancer detection. A linear-SVM model could be used to evaluate the performance of the proposed model. As measured by these criteria outcomes, our model significantly outperforms the SVM model.

Abdar et al., (2019) enhanced an automated ES's breast cancer detection. It first tests SVM parameter settings. These regularization settings improve SVM BC detection, according to research. First-stage accuracy is 99.71%. This study's techniques were evaluated using specificity, sensitivity, and accuracy. The model has excellent accuracy and can increase BC detection algorithm performance.

Wang et. al., (2018) study found that SVMs trained using a variant of the area under the curve (AUC) ensemble learning model were better at breast cancer detection than other methods. The Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) model improves cancer detection. Compare performance variations to emphasise model dependability in sickness diagnosis. On limited datasets, the WAUCE framework may increase diagnostic accuracy by 0.94 percentage points and decrease variance by 69.23%. The WAUCE model outperforms the best single SVM model by 33.34% and 97.19% in accuracy and variance.

Table 1 shows that various authors have used the method and given their findings.

Table 1: Summary of the reviewed literature

Author	Technique	Outcomes
Lanbaran et al., (2021)	Fuzzy-rough set	The selected features show that the approach utilized is more effective.
Chaudhari et al., (2020)	Data Mining	The findings suggest that accuracy increases when variables are limited to those that have been classified as relatively relevant, and the dataset is updated after contradictory variables are removed.
Akinnuwesi et al., (2020)	SVM	Preprocessing breast cancer data, BC-RAED can categorize patients according to risk characteristics and classify instances as either malignant or benign based on established technical indications.
Abdar et al., (2020)	BayesNet Method	The findings outperformed Bayes Algorithm. And it has a very good performance of more than 90% at different values of k.
Mojrian et al., (2020)	Multilayer Fuzzy Expert System of ELM-RBF	Comparing the results to those obtained using SVM, the findings demonstrate high sensitivity, accuracy as well as specificity, true positive rate (TPR), verification, as well as false-negative rate (FNR).
Abdar et al., (2019)	CWW-BANN-SVM Model	The suggested model has 100 percent accuracy and has the potential to increase the performance of algorithms for the detection of BC.
Wang et al., (2018)	AUC ensemble learning algorithm	The suggested WAUCE model, when applied to the SEER breast cancer dataset, increase accuracy by 33.34 percent.

3. Background study

Early medication efficacy prediction is crucial for cancer treatment planning. Using data from neoadjuvant therapy patients, researchers seek to predict breast cancer recurrence within three years (BCR). A new approach was developed using dynamic contrast-enhanced magnetic

resonance imaging (DCE-MRI) pictures before and after therapy. This study analyses and predicts human sickness using Data Mining methods including classification, segmentation, and correlation. In this work, here examined three data mining methods for predicting breast cancer recurrence: the k-nearest neighbor algorithm, Naive Bayes, and decision trees, all of which use principal component analysis (PCA) for feature extraction. The ensemble classifier's top fuzzy features were augmented with clinical information to boost accuracy by 91.7 and 85.2%. (age, risk factor). Medical professionals might use this classifier to predict breast cancer recurrence(Comes et al., 2021).

4. Problem Formulation

Today, breast cancer affects more women than any other disease, but it has also been diagnosed in a few males recently. The fact that they didn't attempt to hide it and instead got the help they needed was the most vital part of this story. Since breast cancer is a potentially fatal disease, early detection is essential for effective treatment. Data mining, support vector machines (SVMs), clustering algorithms, key-value-pair analysis (kNN), and the fuzzy-nearest-neighbor (FNA) test are just some of the methods that have been used to identify it using fuzzy logic. Therefore, these tactics are used using Fuzzy logic for the goal of detection.

5. Research methodology

The focus of this study has been on improving the diagnostic accuracy of a classification method that can distinguish between cancer, benign, and healthy breast parenchyma using mammograms.

Ensemble classifier

To improve a classifier's performance, the ensemble may be employed. Pairing weak and strong students is a helpful meta-categorization strategy that could improve the weak student's performance. In order to improve the precision of several cardiac illness prediction systems, the ensemble approach is used. When many classifiers are combined, they can outperform any one of them alone. In Fig. 3 we can see the ensemble's procedure (Latha et al., 2019).

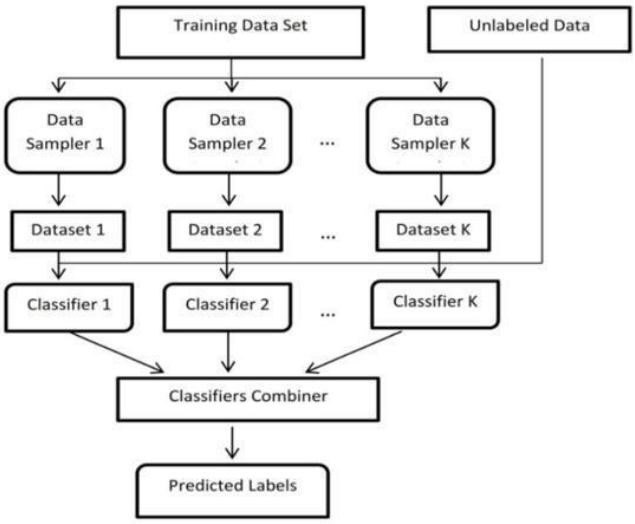

Figure 2: Ensemble process.

6. Proposed Methodology

Dataset 1 and dataset 2 are treated separately in this manner. Following the below-described procedures, dataset 2 is generated from dataset 1, which contains the recurrent patient. Below is a diagram of the suggested methodology's structure, which corresponds to Figure 5.

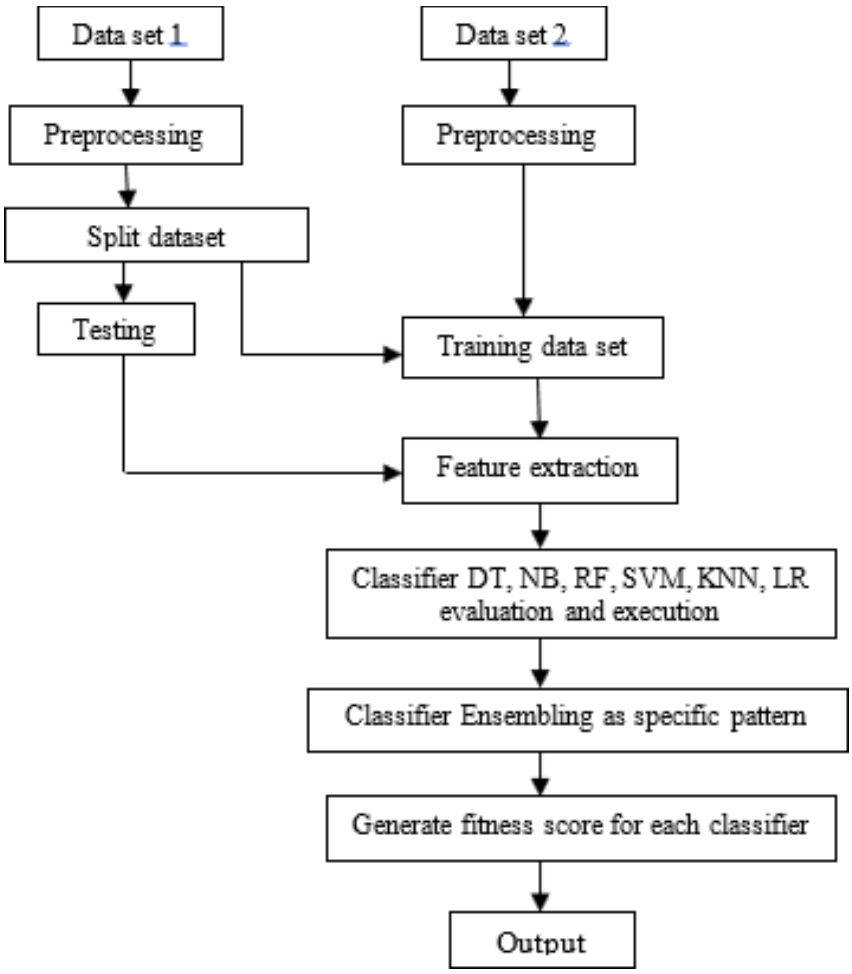

Figure 3 : Proposed Methodology

Step 1: As can be seen in Figure 5, this method considers two datasets, which are labelled as Dataset 1 and Dataset 2 correspondingly. Patients who appear in both datasets are sent to dataset 2, whereas those who appear only in dataset 1 are directed to the current patient.

Step 2: At this stage, have preprocessed both of our data sets (1 and 2) and have separated Data Set 1.

Step 3: Using the age and risk characteristics from both datasets, the fuzzy system is trained to determine which body areas will get cancer again.

Step 4: Feature extraction is often considered the most crucial stage in categorization. Feature extraction is the process of extracting data features for detection and identification. Without losing important data, resources must be reduced. It cleans up the dataset by removing extraneous data.

Step 5: After feature extraction, several classifiers such (DT, NB, RF, SVM, KNN, and LR) are tested and used.

Step 6: At this point, the ensemble classifier is used to describe patterns. In this method, the best fitness score from each classifier (DT&NB, DT&RF) is combined to get an overall fitness score. The best method is the ensemble technique, which pools the ratings of many classifiers to provide more accurate results.

References

Abdar, M., & Makarenkov, V. (2019). CVW-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement*, *146*, 557-570.

Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., & Gururajan, R. (2020). A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognition Letters*, *132*, 123-131.

Akinnuwesi, B. A., Macaulay, B. O., & Aribisala, B. S. (2020). Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. *Informatics in Medicine Unlocked*, *21*, 100459.

Chaudhari, S., Patil, M., & Bambhori, J. (2014). Study and review of fuzzy inference systems for decision making and control. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, *14*(147), 88-92.

Chaudhuri, A. K., Sinha, D., Bhattacharya, K., & Das, A. (2020). An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research. *Int. J. Recent Tech. Eng*, *8*.

Comes, M. C., La Forgia, D., Didonna, V., Fanizzi, A., Giotta, F., Latorre, A., ... & Massafra, R. (2021). Early prediction of breast cancer recurrence for patients treated with neoadjuvant chemotherapy: a transfer learning approach on DCE-MRIs. *Cancers*, *13*(10), 2298.

Ghiasi, M. M., & Zendehboudi, S. (2021). Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, *128*, 104089.

Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques are applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(2), 188-195.

Idris, N. F., & Ismail, M. A. (2021). Breast cancer disease classification using a fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition. *PeerJ Computer Science*, *7*, e427.

Jandoubi, S., Bahri, A., Chakhar, S., & Yacoubi-Ayadi, N. (2015). Mapping the fuzzy semantic model into a fuzzy object-relational database model. In *The seventh international conference on information, process, and knowledge management (eKnow 2015), Lisbon, Portugal, February 22, 2015* (pp. 138-143). International Academy, Research, and Industry Association.

Jevtic, M., Velicki, R., Popovic, M., Cemerlic-Adjic, N., Babovic, S. S., & Velicki, L. (2010). Dietary influence on breast cancer. *J BUON*, *15*(3), 455-461.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

LANBARAN, N. M., & ÇELİK, E. (2021). Prediction of Breast Cancer through Tolerance-based Intuitionistic Fuzzy-rough Set Feature Selection and Artificial Neural Network. *Gazi University Journal of Science*, 1-1.

Lashari, S. A., Ibrahim, R., & Senan, N. (2014, December). De-noising analysis of mammogram images in the wavelet domain using hard and soft thresholding. In *2014 4th World Congress on Information and Communication Technologies (WICT 2014)* (pp. 353-357). IEEE.

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

Magboo, V. P. C., & Magboo, M. S. A. (2021). Machine Learning Classifiers on Breast Cancer Recurrences. *Procedia Computer Science*, *192*, 2742-2752.

Mojrian, S., Pinter, G., Joloudari, J. H., Felde, I., Szabo-Gali, A., Nadai, L., & Mosavi, A. (2020, October). The hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system. In *2020 RIVF*

International Conference on Computing and Communication Technologies (RIVF) (pp. 1-7). IEEE.